REMARKS

Claims 1-48 are pending in the Application. Claims 1-48 have been rejected. The foregoing amendments to the claims are supported at least by original filed Claims 13 and 39, now cancelled. No new matter has been introduced. Acceptance is respectfully requested.

Applicants provide a computer apparatus and method for extracting data from a web page according to the following steps. First, using natural language processing or lexical analysis, possible formal names (organization names and/or people names) are found on a given web page to produce a first set of formal names. Next, the given web page is searched for formal names not found by using natural language processing to produce a second set of formal names. In a preferred embodiment of the present invention, the second set of formal names are found by using pattern matching techniques. Finally, the sets of formal names are combined and refined to produce a working set of people and organization names extracted from the given web page.

Claims 1, 3-5, 7, 9-11, 13-24, 26-31, 33-41, 44-46, and 48 have been rejected under 35 U.S.C. § 102(a) as being anticipated by Paik et al. (U.S. Pat. No. 6,076,088) ("Paik").

The Paik cited reference is directed to an information extraction system that "allows users to ask questions about documents in a database, and responds to queries by returning possibly relevant information which is extracted from the documents." Col. 3, lines 38-41. In particular, the system extracts Concept-Relation-Concept triples (CRCs) from a set of documents and stores them in a database. Col. 3, lines 50-51. The first concept may be a proper name. The system performs similar processing on query text. The system extracts CRCs using natural language processing (NLP) techniques. Col. 4, lines 61-62.

Referring to Fig. 3 of Paik, a block diagram shows the document processing performed by the system prior to CRC extraction. Raw documents are provided to a Document Processing module 100. This module includes a Part-Of-Speech (POS) Tagger 145 that identifies grammatical forms and punctuation marks. An Apposition Identifier 155 of the Document Processing module 100 identifies appositional phrases according to specific linguistic patterns contained in the Apposition Evidence database 157. A Proper Name Interpreter 160 of the Document Processing module 100 first locates the boundaries of proper noun phrases using the POS tags and heuristics (to identify proper name phrases which contain embedded conjunctions

and prepositions, e.g., Centers for Disease Control and Prevention). Once the proper name phrases are identified, they are classified according to 53 concept categories.

Paik presents a classification process in which the proper names are examined to determine the appropriate category for each identified proper name. First, the proper name suffixes, prefixes, and infixes are examined using a Proper Name Prefix/Suffix database 162. The proper names are then passed to a database to determine if an alternative, standard form exists (e.g., President Bill Clinton for Bill Clinton). "If the proper name is an alias, the standard form is used for categorization." Col. 11, lines 58-59. Next, the proper names are compared to a name database 162 of significant personal names. Finally, the proper names are run through numerous heuristic tests, such as a context heuristic test, until the proper names have been tested to fit in one of the 53 categories. As shown in Fig. 3, the Proper Name Interpreter 160 provides its result to a Concept Identifier 165. The proper names identified in Proper Name Identifier 160 are considered as concepts. The concepts identified by the Concept Identifier 165 are submitted to a Sense Disambiguator 170 to assign a unique sense (in the form of a concept) to each content bearing word in the text.

After document processing, the system extracts CRCs using a CRC extractor 105 shown in Fig. 4A. The CRC extractor includes various modules whose outputs are communicated to a CRC combiner 220. The Special Linguistic Construction based CRC Extractor 200 "identifies semantic relations between concepts using the co-referential proper name algorithm and the relation revealing formula." Col. 14, lines 19-21. This module is designed to process text that is redundant. The Special Semantic Relation based CRC Extractor 205 "extracts semantic relations, looking for specialized types of concepts and linguistic clues, including some prepositions, punctuation, or specialized phrases." Col. 16, lines 34-36. CRCs are extracted according to pre-specified rule patterns contained in the Semantic Relation specific CRC Extraction Rule Base 207.

The Syntactic Relation to Semantic Relation Mapper 210 "maps syntactic relations such as 'subject of the transitive verb' to their semantic functional equivalents so that a subject of a verb might be described as 'agent of the action' of a verb." Col. 17, lines 63-66. Finally, the Temporal Information Extractor 215 "time stamps extracted information in order to allow [the system] to present it to the user as part of an automatically constructed time line for any named

entity." Col 19, lines 51-54. "The Temporal Information Extractor 215 extracts information which has a time element." Col. 20, lines 1-2.

The above four CRC extraction modules provide an output to a CRC combiner 220 which removes redundant CRCs extracted by the different modules.

As described above, Paik teaches a Proper Name Interpreter 160 which identifies and classifies proper noun phrases. The Proper Name Interpreter 160 identifies proper noun phrases by using POS Tags and heuristics. Applicants, on the other hand, search for formal names not found by natural language processing using pattern matching techniques to produce a second set of formal names. Paik does use specific linguistic patterns contained in the Apposition Evidence Database to identify appositional phrases. Paik, however, does not teach using pattern matching techniques to find organization names and/or people names. Paik also discloses the use of prespecified rule patterns to interpret the meaning of text and extract CRCs. The prespecified rule patterns, however, are not used to actually identify proper names. In fact, the Special Semantic Relation Base CRC Extractor 205 "relies heavily on the proper name . . . concept categories established during document processing." Col. 16, lines 42-44. The prespecified rule patterns are only used to extract CRCs. The other CRC extraction modules similarly aid in extracting CRCs after proper noun phrases have been identified. In contrast, Applicants use pattern matching techniques to find additional formal names not found using natural language processing techniques.

Independent Claim 1 has now been amended to recite using pattern matching techniques to search a given web page for formal names not found by the natural language processing to produce a second set of formal names. Since Paik does not teach, suggest, or otherwise make obvious each and every claim limitation of now amended independent Claim 1, Applicants respectfully request that the rejection under 35 U.S.C. § 102(a) of independent Claim 1 be withdrawn.

Claim 13 has been cancelled in favor of base Claim 1. Claims 3-5, 7, 9-11, and 14 depend from base Claim 1. Therefore, Applicants respectfully request that the rejection of these claims be withdrawn for at least the same reasons.

Independent claim 15 recites detecting a regular recurrence of certain types of elements to produce additional formal names. In addition to the arguments above with respect to

independent Claim 1, Paik discloses the use of linguistic patterns or prespecified rule patterns at a grammatical level. See column 16, lines 60-67. Applicants instead detect a regular recurrence of certain types of elements throughout a given web page document to produce additional formal names. Applicants have amended claim 15 to clarify this distinction. Since Paik does not teach, suggest, or otherwise make obvious each and every claim limitation of now amended independent Claim 15, Applicants respectfully request that the rejection of this claim be withdrawn.

Claims 16-24, 26-31, and 33-37 depend from base claim 15. Therefore, Applicants respectfully request that the rejection of these claims be withdrawn for at least the same reasons.

Independent claim 38 has now been amended to include all of the limitations of Claim 39 which depends from independent Claim 38. Independent Claim 38 as now amended includes the same limitations as independent Claim 1. Therefore, Applicants respectfully request that the rejection of now amended independent Claim 38 should be withdrawn for at least the same reasons as set forth above with respect to independent Claim 1.

Dependent Claim 39 has been cancelled. As a result, dependent Claims 40-43 have been amended to depend from now amended independent Claim 38 instead of dependent Claim 39.

Claims 40-41, 44-46, and 48 depend from now amended base Claim 38. Therefore, Applicants respectfully request that the rejection of these claims be withdrawn for at least the same reasons given with respect to base Claim 38.

Claims 2 and 25 have been rejected under 35 U.S.C. 103(a) as being unpatentable over Paik in view of Asija (U.S. Pat. No. 4,270,182).

As explained above, Paik does not teach, suggest or otherwise make obvious each and every limitation of now amended independent Claim 1. Asija does not add to Paik the feature of using pattern matching techniques to find formal names not found by natural language processing to produce a second set of formal name as claimed in base Claim 1. Since Claim 2 depends from base Claim 1, Applicants respectfully request that the rejection of this claim be withdrawn for at least the same reasons.

As explained above, Paik does not teach, suggest or otherwise make obvious each and every limitation of now amended independent Claim 15. Asija does not add to Paik the detecting of a regular recurrence of a certain type of element throughout a web page document as claimed

in base Claim 15. Since claim 25 depends from base claim 15, Applicants respectfully request that the rejection of this claim be withdrawn for at least the same reasons.

Claims 6, 8, 32, and 42-43 have been rejected under 35 U.S.C. 103(a) as being unpatentable over Paik as applied to claims 1 and 7 above in view of Brady et al. (U.S. Pat. No. 6,463,430) ("Brady").

As explained above, Paik does not teach, suggest or otherwise make obvious each and every limitation of now amended independent Claims 1 and 38. Brady does not add to Paik the claimed feature of using pattern matching techniques to find formal names not found by natural language processing to produce a second set of formal names. Since claims 6 and 8 depend from base Claim 1 and Claims 42-43 depend from base Claim 38, Applicants respectfully request that the rejection of these claims be withdrawn for at least the same reasons.

As explained above, Paik does not teach, suggest or otherwise make obvious each and every limitation of now amended independent Claim 15. Brady does not add to Paik the claimed detecting a regular recurrence of a certain type of element throughout a web page document. Since Claim 32 depends from base Claim 15, Applicants respectfully request that the rejection of this claim be withdrawn for at least the same reasons.

Claims 12 and 47 have been rejected under 35 U.S.C. 103(a) as being unpatentable over Paik, as applied to claims 1 and 38, in view of Smith et al. (U.S. Pat. No. 6,052,693) ("Smith").

As explained above, Paik does not teach, suggest or otherwise make obvious each and every limitation of now amended independent Claims 1 and 38. Smith does not add to Paik the claimed feature of using pattern matching to find formal names not found using natural language processing to produce a second set of formal names. Since Claims 12 and 47 depend from base Claims 1 and 38, respectively, Applicants respectfully request that the rejection of these claims be withdrawn for at least the same reasons.

CONCLUSION

In view of the above amendments and remarks, it is believed that all claims (Claims 1-12, 14-38 and 40- 48) are in condition for allowance, and it is respectfully requested that the application be passed to issue. If the Examiner feels that a telephone conference would expedite prosecution of this case, the Examiner is invited to call the undersigned.

Respectfully submitted,

HAMILTON, BROOK, SMITH & REYNOLDS, P.C.

By Mary Lou Wakimura

Mary Lou Wakimura
Registration No. 31,804
Telephone: (978) 341-0036
Facsimile: (978) 341-0136

Concord, MA 01742-9133
Dated: 9/26/05